Research on the target customer sales strategy of electric vehicles based on the SVM model of particle swarm optimization

Wen Zhou[1,a,*]
School of Cyber Science and Engineering, Southeast University, Nanjing, China
[a]3095356311@qq.com

Abstract: From the perspective of constructing the classification model, this paper uses the weight coefficient (influencing factors) in the model to analyze the sales impact on different brands of electric vehicles, and optimizes the existing sales strategy. The brands included include joint venture brand 1, independent brand 2 and new power brand 3. The index characteristics of the brand and the personal characteristics of customers are often important factors to determine whether users buy or not. In the process of constructing the classification model, there is the overlapping influence of index features. In order to reduce the influence of collinearity features in the data on the model, the collinearity features of the data are eliminated by using the algorithm based on variance expansion factor. It is obtained that the collinearity feature variables that brand 1 needs to eliminate are A2 (comfort), A4 (safety), A6 (driving operability), B8 (date of birth), B10 (working years) B14 (individual annual income), B15 (family annual income); The collinearity characteristic variables that brand 2 needs to eliminate are A2 (comfort), B8 (year of birth), B10 (working years); The collinearity characteristic variables that brand 3 needs to eliminate are A2 (comfort), B8 (year of birth), B10 (years of work). Then, using Matlab based programming and SVM-RFE algorithm, we get the four influencing factors that have a great impact on the sales results of brand 1, namely B11 (the nature of the user's work unit), A3 (automobile economy), A1 (technical performance of automobile battery) and B9 (the highest level education of the user); The four factors that have a great impact on the sales results of brand 2 are B9 (the highest level education of the user), A1 (technical performance of automobile battery), B11 (nature of the user's work unit) and A5 (performance of automobile power);
The four factors that have a great impact on the sales results of brand 3 are A3 (automobile economy), B14 (individual annual income), B15 (user's family disposable income) and B16 (the proportion of user's annual housing loan expenditure in the total family income). Based on particle swarm optimization algorithm and SVM classification model, aiming at minimizing the cost of improving service difficulty and raising user experience satisfaction to the purchase threshold, select the optimization strategy for customer 1 in brand 1, customer 7 in brand 2 and customer 11 in brand 3. Through the artificial intelligence algorithm, it is concluded that for customer 1, the A1 index should be increased by 0.768%, A3 index by 0.565%, a8 index by 0.007%. If the other indexes do not need to be improved, the user can decide to buy; For customer 7, the A1 index should be increased by 1.549%. If the other indexes do not need to be improved, the user can decide to buy; For customer 11, the A3 index should be increased by 2.795%. If the other indexes do not need to be improved, the user can judge it as purchase.

CSS CONCEPTS • Computing methodologies~Machine learning~Learning paradigms~Supervised learning~Supervised learning by classification

Keywords: SVM-RFE algorithm; Collinearity feature, sales strategy, particle swarm optimization algorithm

I. Data preprocessing

There are 25 characteristic variables in the data, which are A1 battery technical performance and A2 Comfort, A3 economy, A4 safety, A5 power, A6 driving maneuverability, a7 appearance and interior upholstery, a8 configuration and quality, B1 household registration, B2 city residence, B3 living area, B4 driving age, B5 family permanent resident population, B6 marriage and family status, B7 number of children, B8 birth year, B9 highest education, B10 working years, B11 work unit nature, B12 position B13 family annual income (10000), B14 individual annual income (10000), B15 family disposable annual income (10000), B16 annual housing loan expenditure in the proportion of total household income (%), B17 annual car loan expenditure in the proportion of total household income (%).

Due to different feature attributes and value distribution, each feature has different contribution to the classification results. For the classification problem, the difference of each category is related to the selected features, and the collinearity between the features will greatly reduce the stability and accuracy of the regression model. Therefore, in order to reduce the impact of collinearity features on classification problems, it is necessary to eliminate redundant features and retain important features through feature screening. The variance expansion factor algorithm based on multicollinearity diagnosis is often used to eliminate the collinearity characteristics of data. By testing the linear correlation between independent variables, the variance expansion factor algorithm can select independent variables with better independence and enhance the interpretation ability of the model.

The standard definition of variance expansion factor [1] is shown in formula:

$$VIF_i = \frac{1}{1-R_i^2} \tag{1}$$

Where R_i^2 is the sample decision coefficient, expressed as the ith variable X_i correlation with all other variables $X_j(j = 1,2,\dots,k, i \neq j)$, and the size of R_i^2 determines X_i to X_j's ability to explain. If R_i^2 is small, indicating that this factor is interpreted by other factors to a low degree and has a low degree of linear correlation.

Therefore，the higher VIF_i ,the stronger the linear correlation between $X_i$and X_j，When VIF_i is greater than or equal to the threshold，X_i should be eliminated. Variance expansion factor screening criteria : [2]：

1.$VIF_i < 5$， It is considered that there is no collinearity;

2. $5 \leq VIF_i \leq 10$, It is considered that there is moderate collinearity;

$VIF_i > 10$, It is considered that collinearity is serious.

Here, considering the sufficient number of samples of brand 1 and brand 2, set VIF_i threshold is 5. Brand 3 has few samples，so VIF_i threshold is set to 10, so as to retain the characteristic dimension of the data to the greatest extent, so as to enhance the generalization ability of the new power brand (brand 3) model.

The collinearity characteristic variables to be eliminated for the joint venture brand are A2, A4, A6, B8, B10, B14 and B15; The collinearity characteristic variables that independent brands need to eliminate are A2, B8 and B10; The collinearity characteristic variables that need to be eliminated by new power brands are A2, B8 and B10.

II、Analysis of influencing factors of purchase

2.1 Principle and construction of brand purchase intention model based on SVM

Support vector machine is very popular in classification and regression problems. Support vector machine, also known as maximum interval classifier, divides the original sample set into several parts by separating hyperplane.

Now consider a linear binary classification problem for a given sample data(x_1, y_1)、(x_2, y_2)、…、(x_i, y_i)，$x_i \in R^n$ Input data for (user experience data)， $y_i \in \{0,1\}$ Output data for (user purchase intention)， $i =$

$1,2,\ldots,l$.It is hoped that a hyperplane can be found to separate different types of data sets. Considering that most of the real data are not linearly separable, relaxation variables are introduced ξ， establish soft interval support vector machine. That is, the following optimization problems need to be constructed and solved, such as formula:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &st.\, y_i(w^T x_i + b) \geq 1 - \xi_i \\ &\quad \xi_i \geq 0 \end{aligned} \tag{2}$$

Where, C is called the penalty factor, and C > 0, which is used to control the excess ε The degree of punishment of the sample. $\xi_i$is the relaxation factor introduced to ensure that the objective function has a solution under constraints. Since the objective function is a convex function, the optimal solution of the problem is a convex quadratic programming problem. For this objective function, Lagrange function can be defined, as shown in formula

$$L(\alpha,\mu,w,b) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{l}\mu_i\xi_i \tag{3}$$

Among them, α , μ ≥ 0 is a Lagrange multiplier. Under the condition that the kernel function satisfies Mercer, the Lagrange function is also a convex function, and its local optimal solution is also the global optimal solution.In this way, the original objective function of the optimal solution problem becomes a convex quadratic programming problem, as shown in formula :

$$\max_{\alpha,\mu}\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

To $L(\alpha,\mu,w,b)$ calculate w、b、ξ_i the partial derivative of I and make the partial derivative equal to 0 to obtain formulas:

$$\begin{aligned} &\frac{\partial L}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{l}\alpha_i y_i x_i \\ &\frac{\partial L}{\partial b} = 0 \longrightarrow \sum_{i=1}^{l}\alpha_i y_i = 0 \\ &\frac{\partial L}{\partial \xi_i} = 0 \longrightarrow \alpha_i = C - \mu_i \end{aligned} \tag{5}$$

The optimization problem of dual form can be obtained from the above formula, as shown in formula:

$$\begin{aligned} &\max_{\alpha} L(\alpha,\mu,w,b) = \max_{\alpha}\left(\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \left(x_i^T x_j\right)\right) \\ &\quad st. \sum_{i=1}^{l}\alpha_i y_i = 0 \\ &\quad 0 \leq \alpha_i \leq C \end{aligned} \tag{6}$$

Obviously, the optimization problem of dual form is also a quadratic programming problem. The optimal solution can be obtained by using SMO optimization algorithm [3]。 According to KKT conditions, formula (9) can be obtained:

$$\begin{cases} \alpha_i \geq 0 \\ y_i(w^T x_i + b) - 1 + \xi_i \geq 0 \\ \alpha_i[y_i(w^T x_i + b) - 1 + \xi_i] = 0 \\ \mu_i \geq 0 \\ \xi_i \geq 0 \\ \mu_i\xi_i = 0 \end{cases} \tag{7}$$

Therefore, formula this can be obtained:

$$b = y_i - \sum_{j=1}^{m}\alpha_j y_j \left(x_j^T x_i\right), \quad 0 \leq \alpha_i \leq C \tag{8}$$

To sum up, the equation of hyperplane is formula :

$$f(x) = \sum_{i=1}^{m}\alpha_i y_i \left(x_i^T x\right) + b \tag{9}$$

Where m is the number of support vectors. As can be seen from the above formula, only $0 \leq \alpha_i \leq C$， the weight vector value w contributes, and X in the corresponding sample set_ I contributes to the estimation of regression function, so $0 \leq \alpha_i \leq C$ is called support vector. Therefore, the principle of SVM can be represented in Figure1.

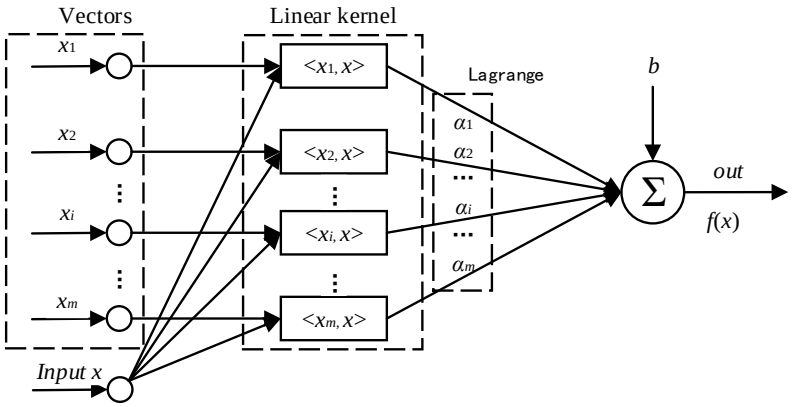

Figure1: principle block diagram of SVM

2.2 Recursive feature elimination feature selection algorithm based on SVM

According to the principle of SVM in the previous section， $w = (w_1, w_2, \ldots, w_m)$是 s the feature weight vector, and for a feature j， w_j larger， means that feature J contains more classification information. Guyon et al. [4] combined SVM with the search strategy of backward recursive elimination and proposed SVM-RFE algorithm. The essence of the algorithm is an encapsulated selection algorithm of heuristic search strategy. By sorting the weight coefficients of each dimension on the SVM hyperplane, deleting the E% features sorted at the end, re modeling the reserved features and sorting the feature weights, and repeating the above operations until only the last feature is left.

The above internal parameter is the proportion e of the number of features deleted in each step, and the number of features deleted each time is the current number of features multiplied by E%. Parameter E greatly affects the computational complexity of SVM-RFE method. The selected features will be less in each iteration, and the load will be better in each iteration. Considering that the solution to this problem is to focus on exploring the impact of influencing factors on the sales of electric vehicles from the perspective of building classification model, rather than focusing on building SVM optimal classification model by looking for the optimal feature subset. Therefore, when e is set to 100, the SVM-RFE method becomes a feature sorting of SVM.

SVM-RFE adopts linear kernel function, and the ranking coefficient can be defined as formula[5-7]:

$$Rank(i) = w_i^2 \tag{10}$$

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{11}$$

Where w_i is the weight vector, and i is the number of iterations.

It can be seen from the above that for brand 1, 18 variables screened are used as input variables; For brand 2, the selected 22 variables are used as input variables; For brand 3, the selected 22 variables are used as input variables. At the same time, the training set and test set are divided in the proportion of 7:3.

Using the established SVM model, the penalty parameter is selected as C = 10, the SVM model is trained, and the test set is input for model test. The results of support vector machine model classification are shown in Table 1.

Table 1 Model accuracy

Various brand models	Training set classification	Test classification accuracy
Brand 1 model	97.9487%	92.2156%
Brand 2 model	93.4758%	94.5026%
Brand 3 model	94.7368%	85.3659%

It can be seen from table 3 that through the comparison of the accuracy of the purchase intention models of the three electric vehicle brands, it can be found that the accuracy of the model test set of the first two brands is higher than that of the model of brand 3, which may be due to the wide gap between the sample data of brand 3 and the sample data of the first two brands. It may also be that the data uniformity is poor, the characteristics are relatively not obvious, and the training effect of the model is relatively poor. The overall accuracy of the three brand recognition models is 94.6%, but the overall accuracy of the three brand recognition models is 92.1%; The accuracy of training set and test set of brand 2 was 93.4758% and 94.5026%, respectively; The accuracy of the training set of brand 3 was 94.7368%, and that of the test set was 85.3659%.

2.3 Analysis of influencing factors based on SVM-RFE model

Since the model can better classify the purchase intention of the three brands, the relative importance of the feature variables that affect the classification effect (the importance of the feature that has the greatest impact on the classification effect is 100%, and the relative importance proportion of other features is calculated) is drawn according to the results of SVM-RFE algorithm, as shown in figure.2, figure. 3 and figure.4 respectively.

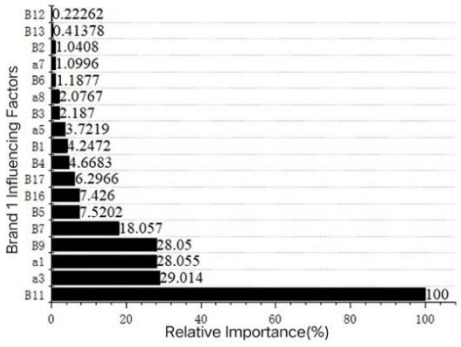

Figure2: Relative importance of brand 1

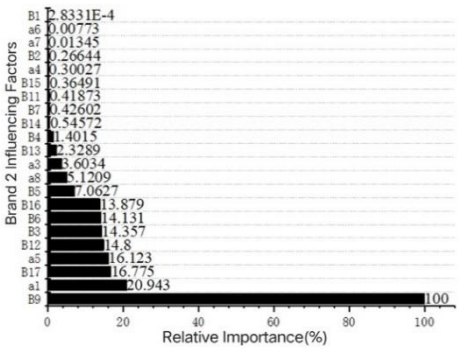

Figure 3: Relative importance of brand 2

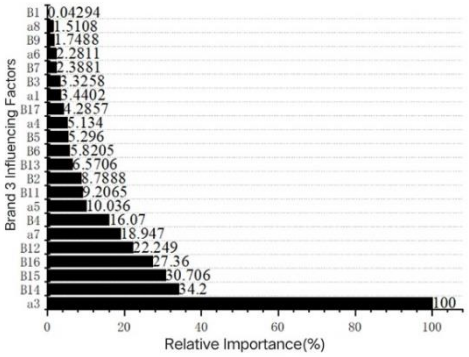

Figure4: Relative importance of brand 3

As can be seen from Figure 2, figure3 and figure 4, the four influencing factors that have a great impact on the classification results of brand 1 are B11 (the nature of the user's work unit), A3 (automobile economy), A1 (technical performance of automobile battery) and B9 (the highest level education of the user), and their relative contributions are more than 25%; The four influencing factors that have a great impact on the results of brand 2 classification are B9 (the highest level education of users), A1 (automobile comfort), B11 (the nature of users' work unit) and A5 (automobile power performance), and their relative contributions are more than 15%; The relative contribution of car loans to the total household income (B13) and B15 (B15) are the factors that affect the annual disposable income of users, accounting for 25% of the total household income respectively;

III、Sales strategy optimization

As we all know, the difficulty of service is directly proportional to the percentage of satisfaction. That is, in order to improve the user's experience satisfaction, the difficulty of service will gradually increase. Therefore, in order to increase the user experience satisfaction to the purchase threshold and minimize the cost of improving the difficulty of the service, the data information of three non purchased users is selected, and the classification models of three brands trained before are used as the discrimination model of user purchase intention. At the same time, the characteristics of a1-a8 are optimized based on particle swarm optimization algorithm, Get the sales strategy result that minimizes the cost of improving the difficulty of service, that is, the satisfaction improvement result in a1-a8. The data information of users who do not buy the corresponding brand here is shown in Table 1.

3.1 Principle and construction of target customer sales strategy model based on particle swarm optimization algorithm and SVM model

3.1.1 Basic principle of particle swarm optimization

Particle swarm optimization (PSO) is a heuristic optimization algorithm based on swarm intelligence. It simulates the behavior of "personal cognition and social cognition" in the process of bird flock foraging, and continuously updates the position of bird flock by updating the flight speed of bird flock. In the implementation of particle swarm optimization program, the bird swarm is simplified into particle swarm optimization, mainly including the key technologies of algorithm model establishment, particle velocity, particle position and inertia weight update.

The mathematical model of the optimization problem corresponds to the particle swarm optimization algorithm model. The optimization variables correspond to the particle position, the objective function corresponds to the particle fitness, and the constraints are reflected in the particle position generation and particle fitness. Optimization is the process of constantly updating the particle flight speed and position.

1. Particle position

In the PSO algorithm model, the particle position is the optimization variable, and the number of optimization variables is the dimension of the particle position. In the process of foraging, the particle position is updated by constantly updating the particle flight speed. The position update formula is shown in :

$$x_{id}^{k} = x_{id}^{k-1} + v_{id}^{k} \ \ (i = 1, \dots, n+1) \tag{12}$$

2. Particle velocity

The direction and amplitude of particle position change depend on the particle flight speed, which affects the optimization effect and convergence speed of the algorithm. It is composed of inertial cognition, self cognition and social cognition. The speed update formula is shown :

$$v_{id}^{k} = w \cdot v_{id}^{k-1} + c_1 r_1 \left(p_{id} - x_{id}^{k-1}\right) + c_2 r_2 \left(p_{gd} - x_{id}^{k-1}\right) \tag{13}$$

The overall process is shown in Figure 5.

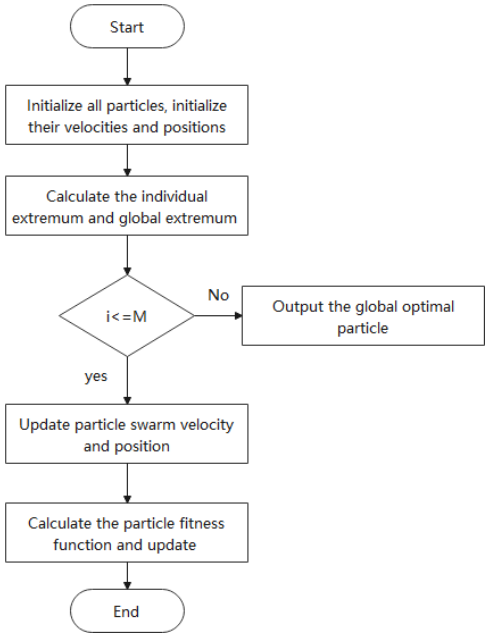

Figure 5: Iterative flow of particle swarm optimization algorithm

3.1.2 Model building

According to preamble, the features retained by brand 1 in a1-a8 features are A1, A3, A5, a7 and A8; The features retained by brand 2 in a1-a8 features are A1, A3, A4, A5, A6, a7 and A8; The features retained by brand 3 in a1-a8 features are A1, A3, A4, A5, A6, a7 and A8. Thus, the optimization mathematical model of problem 4 is established, as shown in formula:

$$\begin{aligned} &\min f_{obj} = \textstyle\sum_{d=1}^{q} \left(x_{id}^{k} - \eta_d\right) \\ &st. \ \textstyle\sum_{d=1}^{q} x_{id}^{k} w_d + \sum_{e=q+1}^{m} \theta_e w_e + b > 0 \\ &\eta_d < x_{id}^{k} < \eta_d + 5 \end{aligned} \tag{14}$$

Among them, η It is the feature experience data retained by the corresponding user in the corresponding brand after removing the collinearity feature from the a1-a8 feature; Q is the number of features retained by the corresponding brand in a1-a8 features; W and B are the weight coefficients and deviation coefficients of the hyperplane of the SVM model of the corresponding brand.

According to the above formula, particle x_i^k optimizes the feature space reserved by the corresponding brand, and minimizes the improvement of user satisfaction (i.e. minimizing the improvement of service difficulty) under

the condition of meeting the constraints. After continuous iteration, it finally finds out the most appropriate user satisfaction and makes users buy cars of the corresponding brand, that is, it formulates the optimal sales strategy.

3.1.3 Solution of target customer sales strategy model based on particle swarm optimization algorithm and SVM model

Set the parameters of PSO model as, the fitness iteration process in the optimization process is shown in Figure 6, figure 7 and figure 8.

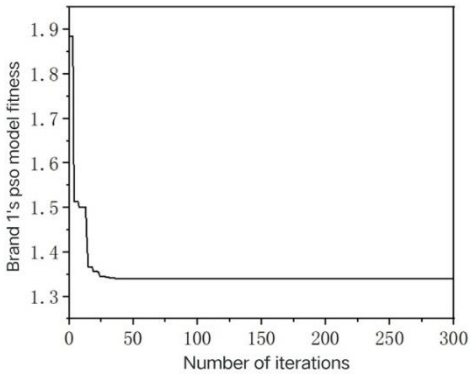

Figure 6: Iterative convergence diagram of PSO model fitness of brand 1

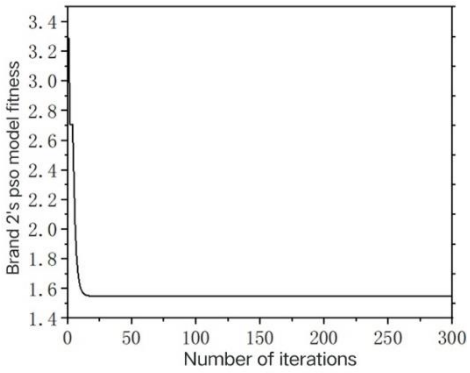

Figure 7: Iterative convergence diagram of PSO model fitness of brand 2

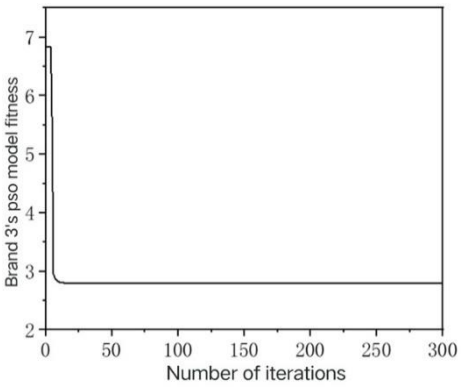

Figure 8: Iterative convergence diagram of PSO model fitness of brand 3

The user satisfaction of the corresponding brand output after optimization iteration is shown in Table 2, table 3 and table 4.

Table 2 customer 1 satisfaction comparison after optimization

	a1	a3	a5	a7	a8
Original user satisfaction	89.844	88.919	88.87	99.993	99.98
Optimized user satisfaction	90.612	89.484	88.87	99.993	99.987

Table 3 customer 7 satisfaction comparison after optimization

	a1	a3	a5	a7	a8
Original user satisfaction	87.074	81.558	85.19	82.614	83.596
Optimized user satisfaction	88.623	81.558	85.19	82.614	83.596

Table 4 customer11 satisfaction comparison after optimization

	a1	a3	a5	a7	a8
Original user satisfaction	88.796	81.558	81.031	80.839	75.027
Optimized user satisfaction	88.796	84.353	81.031	80.839	75.027

IV. Conclusions

This paper uses SVM-RFE algorithm to analyze the influencing factors of car purchase of different brands, and optimizes the sales strategy for specific customers by using the combination of particle swarm optimization algorithm and SVM algorithm. Integrating particle swarm optimization algorithm into SVM model can not only solve high-dimensional problems, but also deal with the interaction of nonlinear features. Finally, it is concluded that for brand 1 joint venture brand cars, people from state-owned enterprises, private enterprises, foreign-funded enterprises and joint ventures buy the most, and the relevant sales departments should sell more to the people of these enterprises, especially the middle-level managers, middle-level technicians and small company owners, with the highest sales success rate, and strive to improve the economy of cars and the technical performance of car batteries; For the independent brand cars of brand 2, the people who pay attention to are still people from state-owned enterprises, private enterprises, foreign-funded enterprises and joint ventures, especially those with college or bachelor's degree and whose housing loan accounts for no more than 10% of their income, and strive to improve the comfort of cars and the performance of batteries; For the new power brand of brand 3, the target customers are generally individuals with an annual income of more than 12W and families with an expenditure of more than 13W. These people are often relatively affluent people. At the same time, efforts should be made to improve the car economy, driving operability, appearance and other factors, these people need to pay more attention to these aspects.

References

[1] Prabaharan, M.; Selvalakshmi, M. Customers Interest in Buying an Electric Car: An Analysis of the Indian Market[C]. IFIP Advances in Information and Communication Technology, 2020, 617: 493-509

[2] Moradkhani, Alireza ; Baharvandi.Analyzing the microstructures of W-ZrC composites fabricated through reaction sintering and determining their fracture toughness values by using the SENB and VIF methods[J]. Engineering Fracture Mechanics,2018,189:501-513

[3] Pchelintsev, Evgeny ; Pergamenshchikov, Efficient estimation methods for non-Gaussian regression models in continuous time [J].Annals of the Institute of Statistical Mathematics,2022,74(1): 113-142

[4] Vinothkumar, S. ; Varadhaganapathy, S.;Ramalingam. Fake News Detection Using SVM Algorithm in Machine Learning[C]. International Conference on Computer Communication and Informatics, ICCCI 2022

[5] Kayadelen, C.; Altay, G.; Sequential minimal optimization for local scour around bridge piers. [J].Marine Georesources and Geotechnology,2022,40(4):462-472

[6] Song, Zhibin; Liu, Shurong; Jiang, Mingyue. Parameter Determination Method of Soil Constitutive Model Based on Machine Learning [J].Wireless Communications and Mobile Computing.2022,2022:11-20

[7] Zhong, Zhicheng ; Gao, Qichen; Zhang, Fudong. Research on classification method of abnormal vibration of pipeline based on SVM [C]. The International Society for Optical Engineering,2022,12169